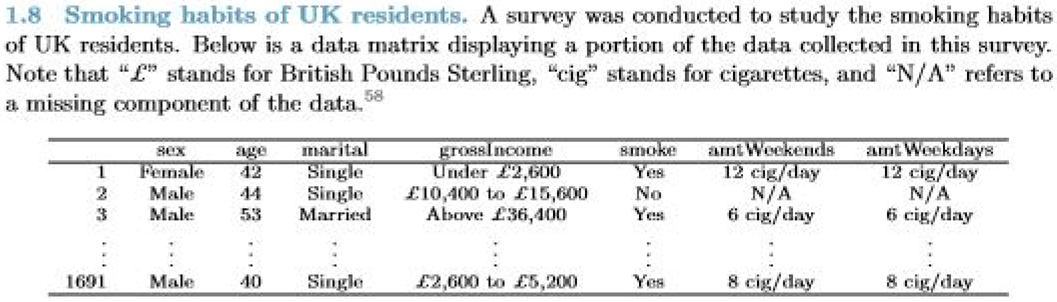
Week_1_Homework

Christophe

August 29, 2015

Chapter 1: 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70



1.8 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[58]

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?

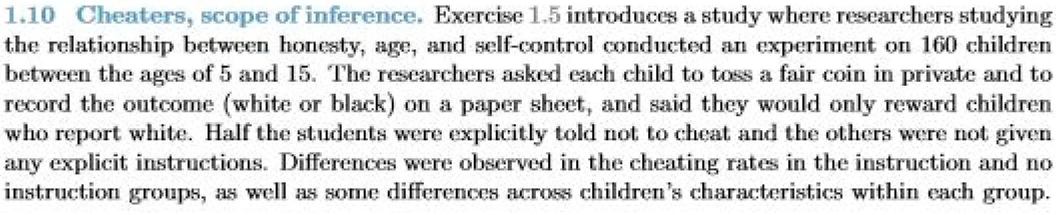
Each row represents an observation with the unique characteristics provided in each column of the data set.

(b) How many participants were included in the survey?

There were 1691 participants

(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

1. sex is categorical and is nominal
2. age is numerical and is techinically continuous but appears to be discrete in this data set
3. martial is categorial and is nominal
4. gross income is numeric and is continuous
5. smoke is categorial and is nominal
6. amtWeekends is categorical and is ordinal
7. amtWeekdays is categorical and is ordinal



1.10 Cheaters, scope of inference. Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

(a) Identify the population of interest and the sample in this study.

The population of interest is children and the sample is 160 children between the ages of 5 and 15.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

The results of the study can generalized to the population because the sample appears to be representative of the true population. However, there could be some unanticipated bias introduced in the sampling process.

Additional tests would be important to confirm the statistical significance of the findings and correlation does not always equal causation. However, because they conducted a randomized experiement they could infer caustion in this study.

1.28 Reading the paper. Below are excerpts from two articles published in the NY Times:

(a) An article titled Risks: Smokers Found More Prone to Dementia states the following:

"Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-aday smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

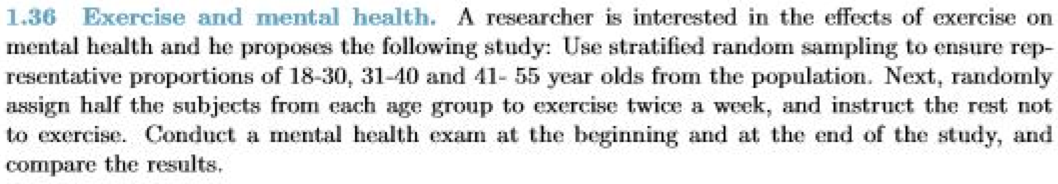
We cannot conclude that there is caustion and that smoking causes dementia. We can only infer that smoking and dementia appear highly correlated and are associated with each other. The reason we are limited in our inferance is that this is simply a health survey and no experiement has taken place to control for other confounding factors. Also, this observational study can not be replicated to confirm results but it does point to a specific area for further studies.

(b) Another article titled The School Bully Is Sleepy states the following:

"The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders." A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children."

Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

The statement is not justified because this is simply an observational study. The conclusion that can be drawn from this article is that sleep disorder and bullying are highly associated but neither can be labeled as causing the other. The friend could just as easily conclude that bullying causes sleep disorders because there is no controlling for confounding issues. The article mentions behavioral issues, which could have a greater impact on the sleep disorder and bullying. Again, the issue is that an experiment would need to take place to determine the cause and effect of these variables and if there were some other factor at play in causing this high association or does one truly cause the other.



1.36 Exercise and mental health. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?

This is a randomized experiment

(b) What are the treatment and control groups in this study?

The treatment group is the the group excersising twice a week and the control group is the group that does not exercise

(c) Does this study make use of blocking? If so, what is the blocking variable?

It does make use of blocking through the age strata, the blocking variable is age.

(d) Does this study make use of blinding?

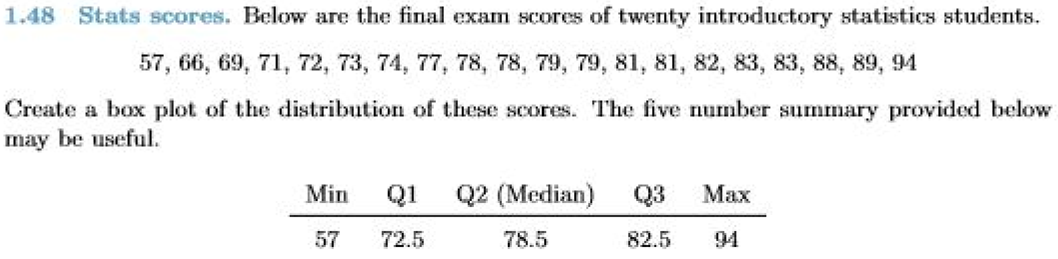
No it doesn't, the patient will know they are in the group that does not exercise as they would be instructed to do as such

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

Only using the information provided the study could be used to establish a causal relationship between exercise and mental health because the use of control groups will allow for testing of only the vairable exercise on the outcome of interest which is mental health. The conclusions could not be used for the population at large because people that are aged below 18 or above 55 are not represented in this study.

(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

I would assume that proposal of this type would have more details but assuming that I do not have more information I would not fund this study as is. The set up appears strong but the outcome of interest is vague as "mental health" can mean a number of things and there is no clear measure. Also, instructing people to not exercise crosses into a moral dilemma and I would have strong reservations about possible enforcing some bad behaviors and even helping a person establish a bad habit of not exercising. I think this could be changed if the participants did not exercise already but then the sample would be skewed to those who do not excerise.



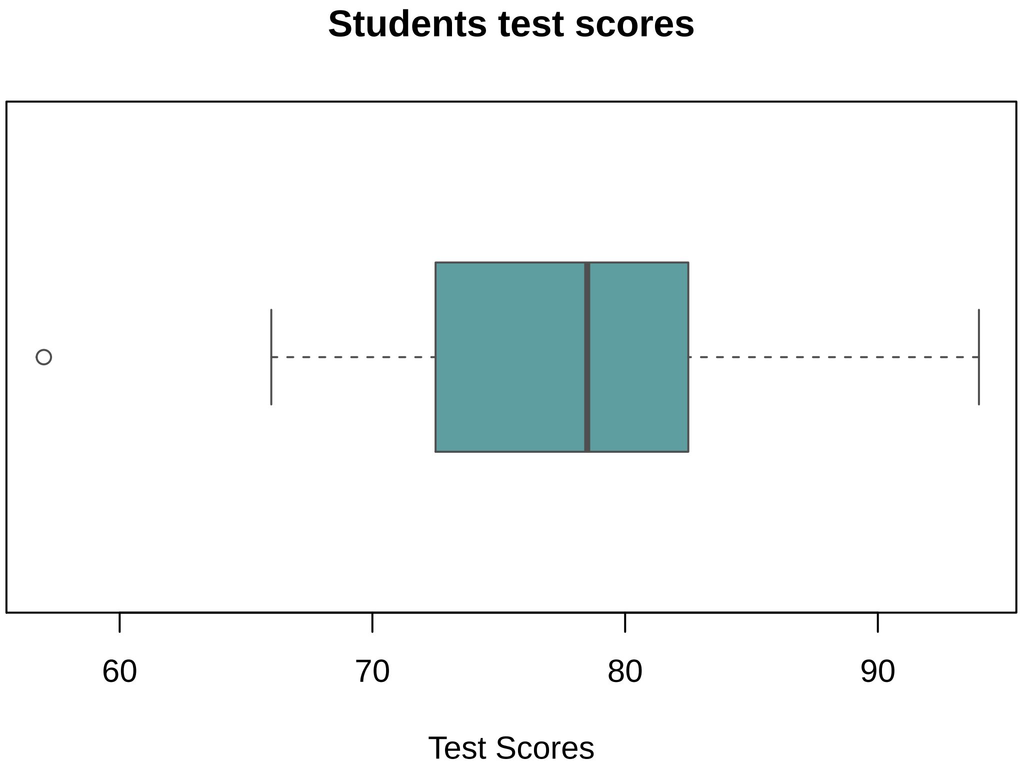
1.48 Stats scores. Below are the final exam scores of twenty introductory statistics students.

57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

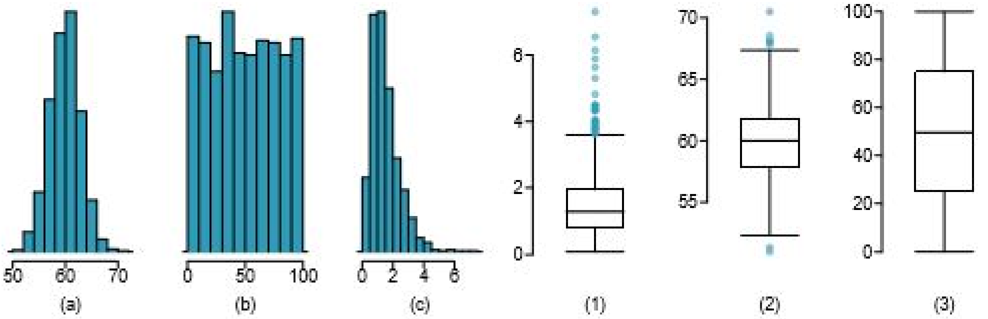
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

```r
scores <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
boxplot(scores, col = "cadetblue", horizontal = TRUE,
        border = "gray31", xlab = "Test Scores", main = "Students test scores")
```



1.50 Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.



(a) is unimodal and is a symetric normal distribution. It is represented by the boxplot of (2)

(b) is multimodal and is represented by the boxplot of (3)

(c) is unimodal and is a right skew distribution. It is represented by the boxplot of (1)

1.56 Distributions and appropriate statistics, Part II . For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below \$350,000, 50% of the houses cost below \$450,000, 75% of the houses cost below \$1,000,000 and there are a meaningful number of houses that cost more than \$6,000,000.

I would expect the distribution to be right skewed. The median would best respresent house prices in this country because the over 6 million dollar houses would skew a mean. The variation would be best explained by IQR, as the over 6 million dollar houses would skew the standard deviation greatly.

(b) Housing prices in a country where 25% of the houses cost below \$300,000, 50% of the houses cost below \$600,000, 75% of the houses cost below \$900,000 and very few houses that cost more than \$1,200,000.

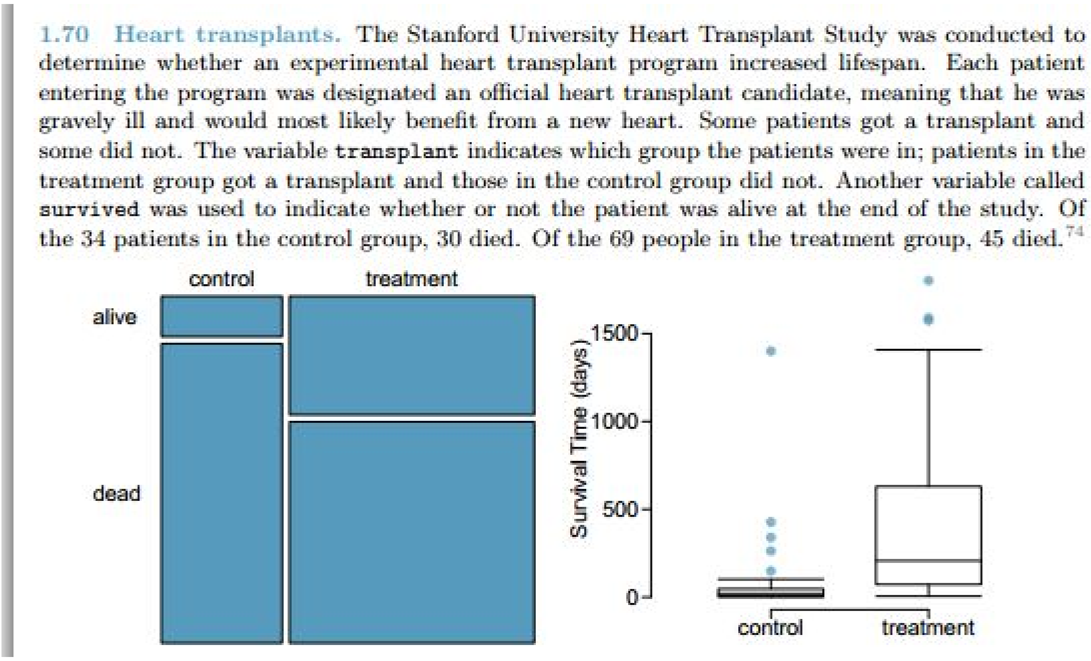
I would expected the distribution to be symetrical and normally distributed. While the mean could potentially provided an acceptable measure of a typical value, it can be changed greatly by any outliers. Therefore, the median remains the best measure of central tendency. The variability of the observation could easily be represented by standard deviation because the distribution appears to be very normal.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

I would assume that this would be right skewed distribution in that a majority of students will drink none or very little. While the mean could potentially provided an acceptable measure of a typical value, it can be changed greatly by any outliers. Therefore, the median remains the best measure of central tendency. The variation would be best explained by IQR, as those that drink excessively may skew a standard deviation greatly.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees

I would assume that this would be a normal distribution of salaries with a vew outliers. Due to the employees that have much greater salaries than the majority of other employees it would be best to use the median as the mean has a breakdown point of 0 so it would be skewed greatly by the outliers. The variation would be best explained by IQR, as those that make significantly more than other employees would be skewing the results.



1.70 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable `transplant` indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called `survived` was used to indicate whether or not the patient was alive at the end of the study. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died.[74]

(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

Survival is not independent of treatment, if it were I would expect relatively similar numbers for control and treatment group. However, we do see a difference in outcomes based on the mosiac plot.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

The boxplot shows a very promising outcome of higher life expectancy for those recieving the treatment than those that did not recieve the treatment. Its fairly intuitive that we would expect gravely ill patients receiving a new heart to at least have a longer life expectancy than those that continue to be gravely ill with no intervention.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

The proportion of patients in the treatment group that died is 45/69 people or 75% and the proportion of patients in the control group that died was 30/34 people or 88%.

(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

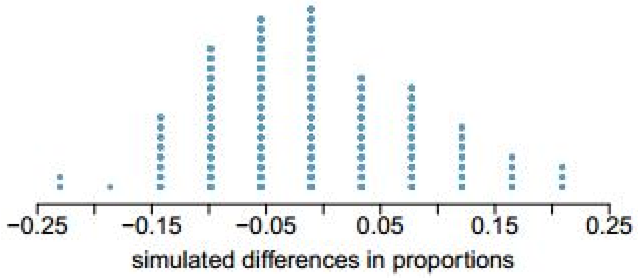
i. What are the claims being tested?

The claim being tested is “does heart transplants affect positive outcomes in patients and extend both survival rates and the length of survivla in patients”

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write alive on 28 cards representing patients who were alive at the end of the study, and dead on 75 cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size 69 representing treatment, and another group of size 34 representing control. We calculate the difference between the proportion of dead cards in the treatment and control groups (treatment -control) and record this value. We repeat this 100 times to build a distribution centered at approximately zero assuming the independence model is true. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are equal to or greater than our observation in the study. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?



The simulation results show that the effectiveness of the transplant program is that transplants has an effect on survival rates. The high rates of survival is actually due to the transplant and not due to random chance.